## Case Study: Ames Iowa Housing Prices

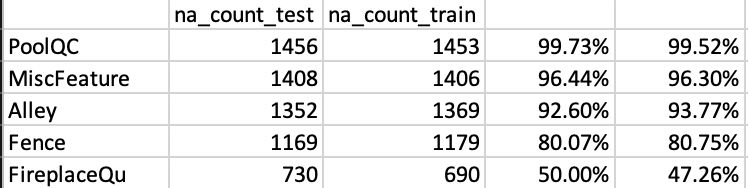
#### Data Description:

See appendix A for a detailed description of the dataset

Data Questions:

Is the client only concerned with residential sales? Should model scope be limited?

#### Missing Data:



Missing data seems to be an issue for a few of the features – namely PoolQC, MiscFeature?

#### Train/Test Balance:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Count | | | Average Square Footage | | |
| Neighborhood | **Average Sale Price** | **Train** | **Test** | **Balance** | **Train** | **Test** | **Balance** |
| BrkSide | 124,834 | 58 | 50 | 86% | 1,203 | 1,272 | 106% |
| NAmes | 145,847 | 225 | 218 | 97% | 1,310 | 1,273 | 97% |
| Edwards | 128,220 | 100 | 94 | 94% | 1,340 | 1,335 | 100% |

Data seems well balanced by neighborhood with respect to the explanatory variables in QOI #1 (GrLivArea and Neighborhood)

Ultimately, training the model will be done on a 5 fold cross validation scheme and then the model will be run against a test set. More analysis on balance should be performed when assessing the full model.

#### Problem Statement:

Century 21 seeks to answer two questions related to housing prices in Iowa:

* QOI 1: Is there a relationship between square footage and sale price?
* QOI 2: Is this relationship dependent on neighborhood?

The data set we will be using will only consider a subset of the available features we could use to build a better model. Since the analysis at hand only seeks to answer questions related to house size and location – we will limit the analysis to only those variables and leave the investigation of further explanatory variable dependencies for future studies.

#### Modeling:

#### QOI: Is there a relationship between square footage (GrLivArea) and sale price (SalePrice)

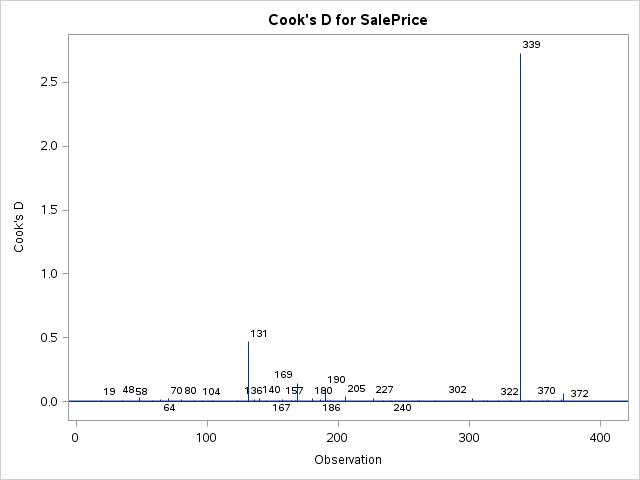
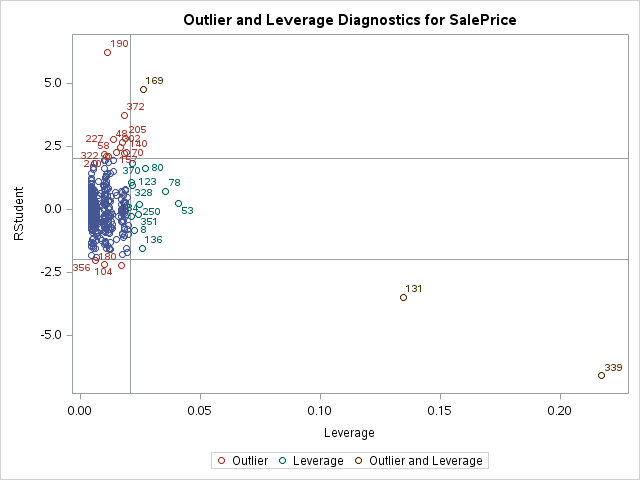
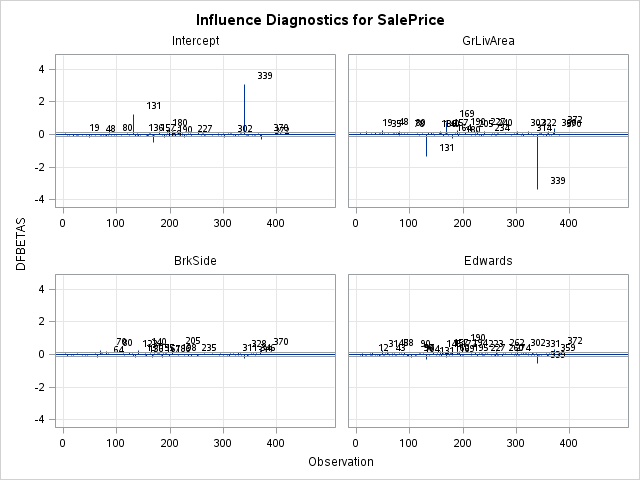
Running a basic linear regression of the form:

SalePrice = 0 + 1 GrLivArea

#### Checking Assumptions / Outliers

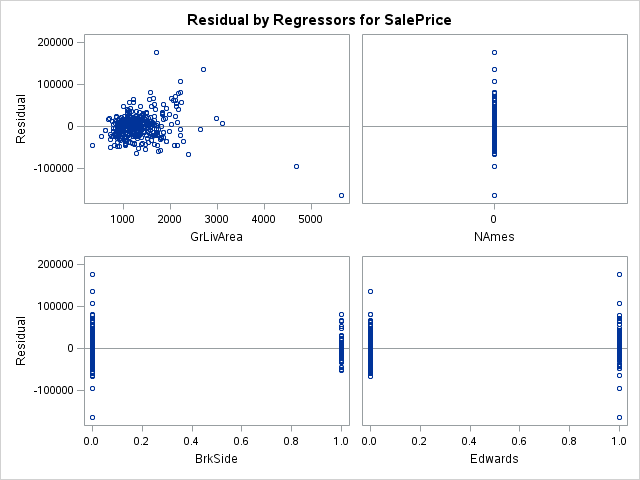
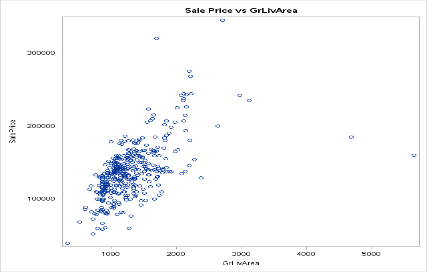
#### Outliers

For the reduced data set (3 neighborhoods) – observations (131, 339) appear to have high leverage/Cooks’ Distance, final model will be run with and without them to assess true influence. Of note – these outliers are most likely to influence according to the differential beta table.

#### Normality:

Residuals generally appear linear and normally distributed amongst population and subgroups

Scatter, Q-Q plot and histogram of residuals are all relatively normal.

#### Linear Trend

The mean sale price as a function of GrLivArea does appear to be relatively linear based on the scatter plot above (increasing GrLivArea tends to have a positive effect on the mean sale price)

#### Equal SD

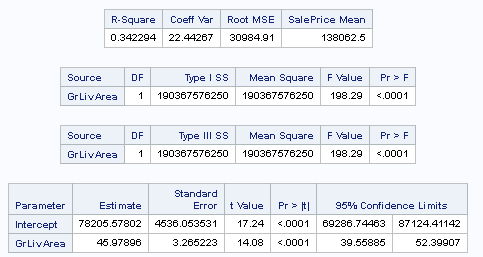
Within the prediction band, the SD of prices appears relatively constant (few signs of heteroscedasticity)

#### Independence

No variables are assumed to be interdependent in this example.

**Model assumptions for basic linear regression seem to be well met. We will proceed with the analysis as is (inference on means)**

#### Simple Linear Model Results



Assuming the only variable that contributes to SalePrice to be GrLivArea – there appears to be strong evidence of a non-zero intercept (0 = {69,286 – 87,124}). In this instance the intercept can be thought of as the cost of the land (i.e a house with no square footage).

A strong direct relationship between GrLivArea and SalePrice also exists, accounting for an effect of 45.9$/sqft of living space – or to state it more clearly – an increase of approx 4,600$ in sale price can be expected from every increase of 100sqft of living space. A 95% CL puts this accretive affect at anywhere from 3,900/100sqft to 5,200/100sqft.

Given the low R-square value for this regression (0.34) it is clear that there are other variables which contribute to SalePrice. GrLivArea only accounts for about 34% of the variance in SalePrice.

#### QOI 2: Does relationship between square footage (GrLivArea) and sale price (SalePrice) vary based on neighborhood?

To address, we will consider a regression of the form (including interaction terms)

SalePrice = 0 + 1 GrLivArea + 2 BrkSide + 3 Edwards + 4 GrLivArea|Edwards + 5 GrLivArea|BrkSide

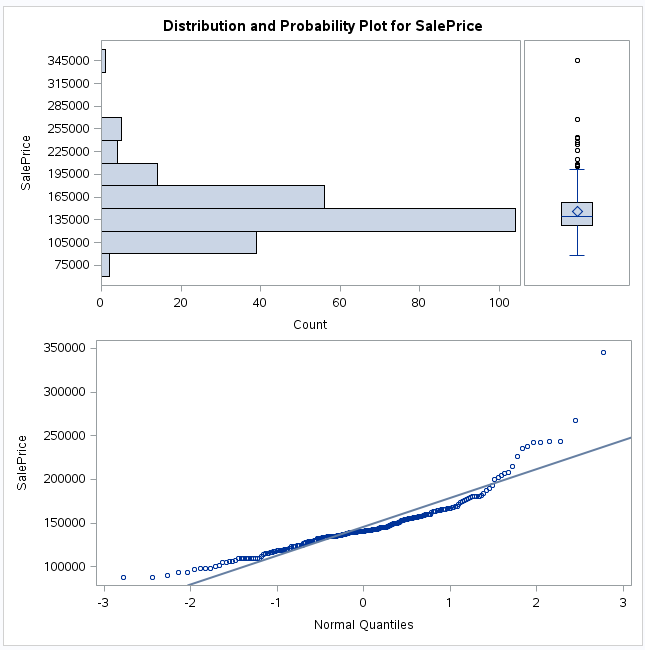
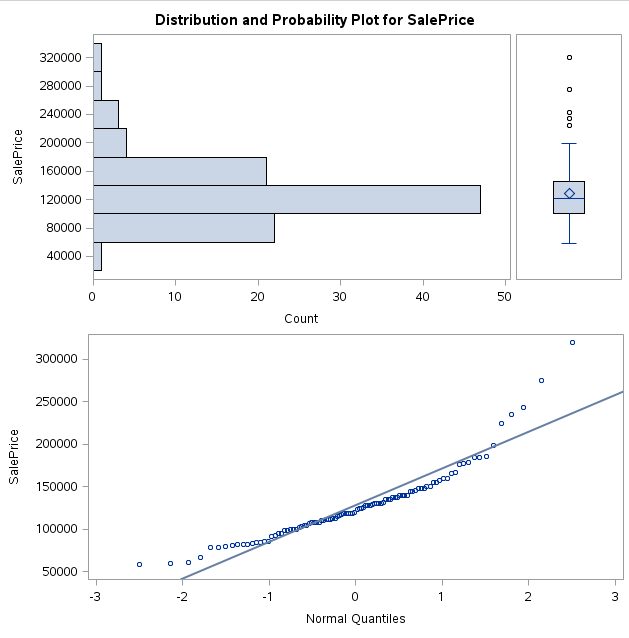
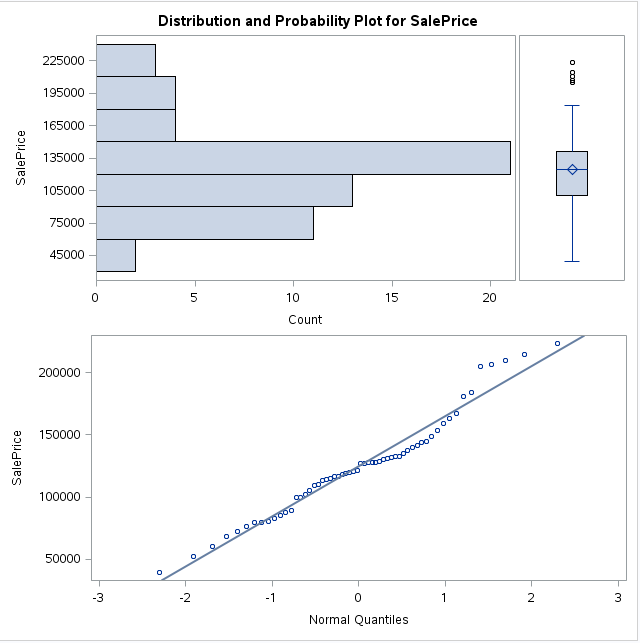
\*reference level for categorical variable is NAmes

#### Revisiting Assumptions:

#### Normality

See QOI 1 for more details – all assumptions hold here

#### BrkSide/Edwards/NAmes



BrkSide shows strong evidence of positive skew though not enough to violate the assumption of normality.

Edwards also shows strong evidence of positive skew though not enough to violate the assumption of normality.

Of the three neighborhoods, Names exhibits the most positive skew though not enough to violate the assumption of normality.

#### Linear Trend

Generally speaking – each of the neighborhoods exhibits a linear relationship between GrLivArea and SalePrice. Edwards being somewhat suspect in that regard having observations with very high square footages and very low sale prices.

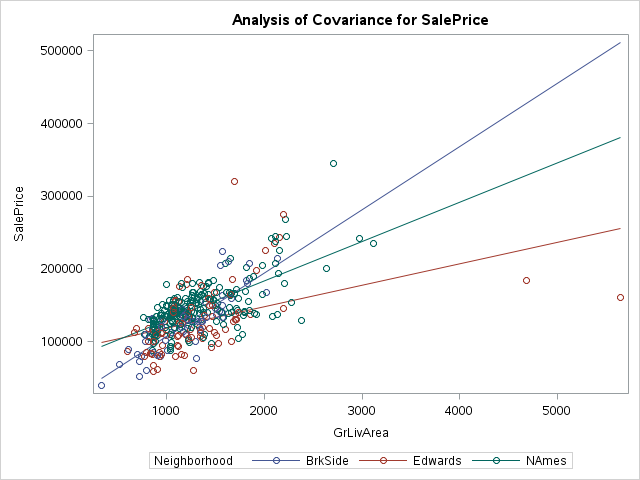
#### Equal SD

Scatter plots do not show any sign of heteroscedasticity amongst the 3 neighborhoods.

#### Independence

Since being in one neighborhood implicitly excludes a home from being in the others – we should be fine to assume independence.

#### Linear Model With Neighborhood Results



#### Interpretation:

All terms show statistical significance – despite the p-value for 2 being greater than our significance level (alpha = 0.01), we will include it in the model since the interaction term with GrLivArea is significant.

#### NAmes:

At the reference level we assume all i = 0 for i >1 which reduces the regression equation to the standard linear model with an expected mean of approximately 74k [62k, 87k] with an expected increase in value of approximately 5,413$ [4,520 – 6,338] per 100sqft of GrLivArea

SalePrice = 0 + 1 GrLivArea : (SalePrice = 74,676+54.31\*GrLivArea)

#### BrkSide:

At the next level we assume all i = 0 for i = 3,5 which reduces the regression equation as follows:

SalePrice = 0 + 1 GrLivArea + 2 BrkSide + 4 GrLivArea|BrkSide : (SalePrice = 19,972+87.16\*GrLivArea)

An expected mean sale price of approximately 20k with an expected increase in value of approximately 8,716$ per 100sqft of GrLivArea

#### Edwards:

At the next level we assume all i = 0 for i =3 which reduces the regression equation as follows:

SalePrice = 0 + 1 GrLivArea + 3 Edwards + 5 GrLivArea|Edwards : (SalePrice = 88,353 + 28.75\*GrLivArea)

An expected mean sale price of approximately 88,353k with an expected increase in value of approximately 2,875$ per 100sqft of GrLivArea

To get an intuitive understanding of these relationships – a home with median square footage (1,464sqft) – would be priced at the following by neighborhood: 

Kaggle Comp

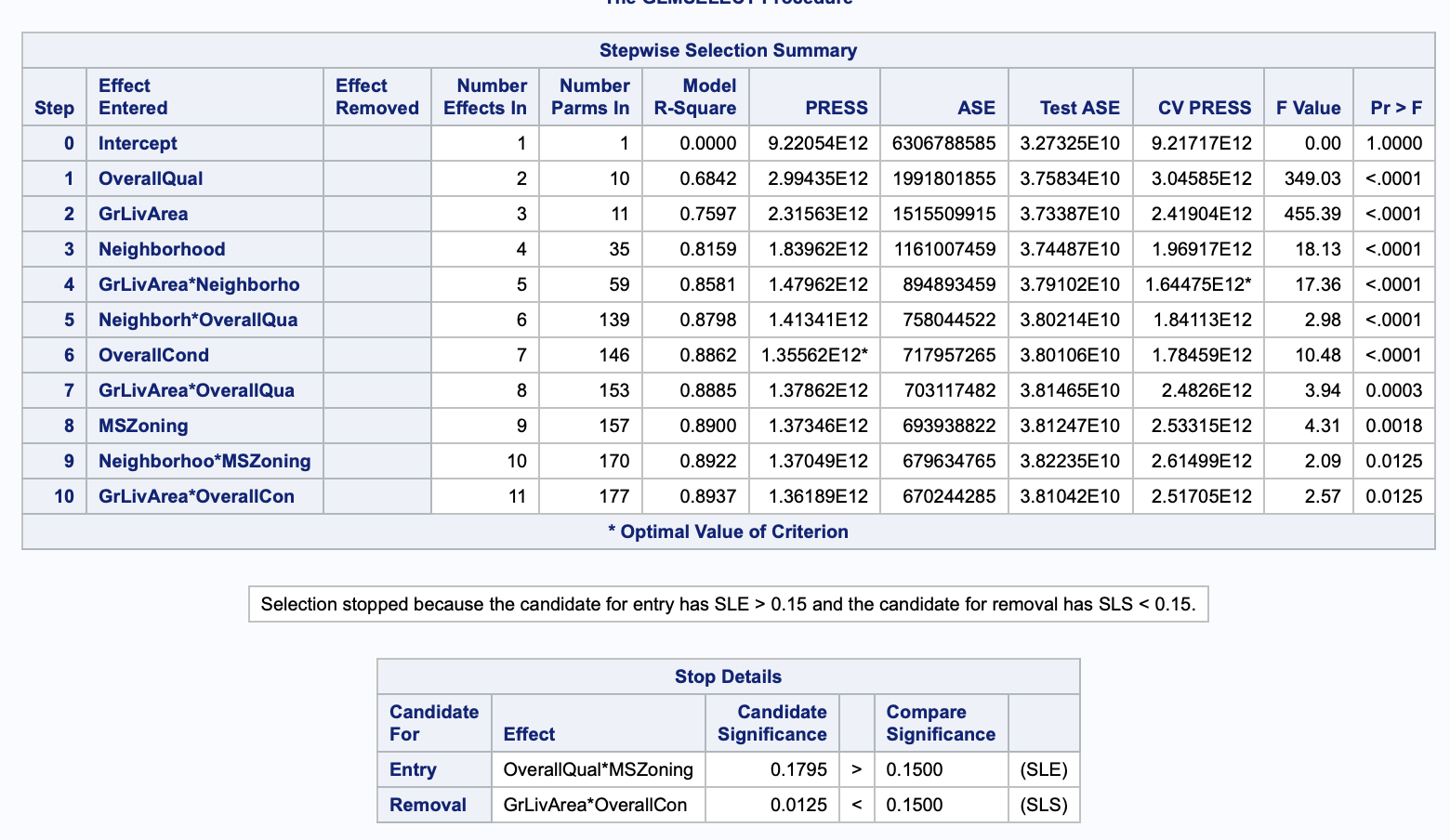
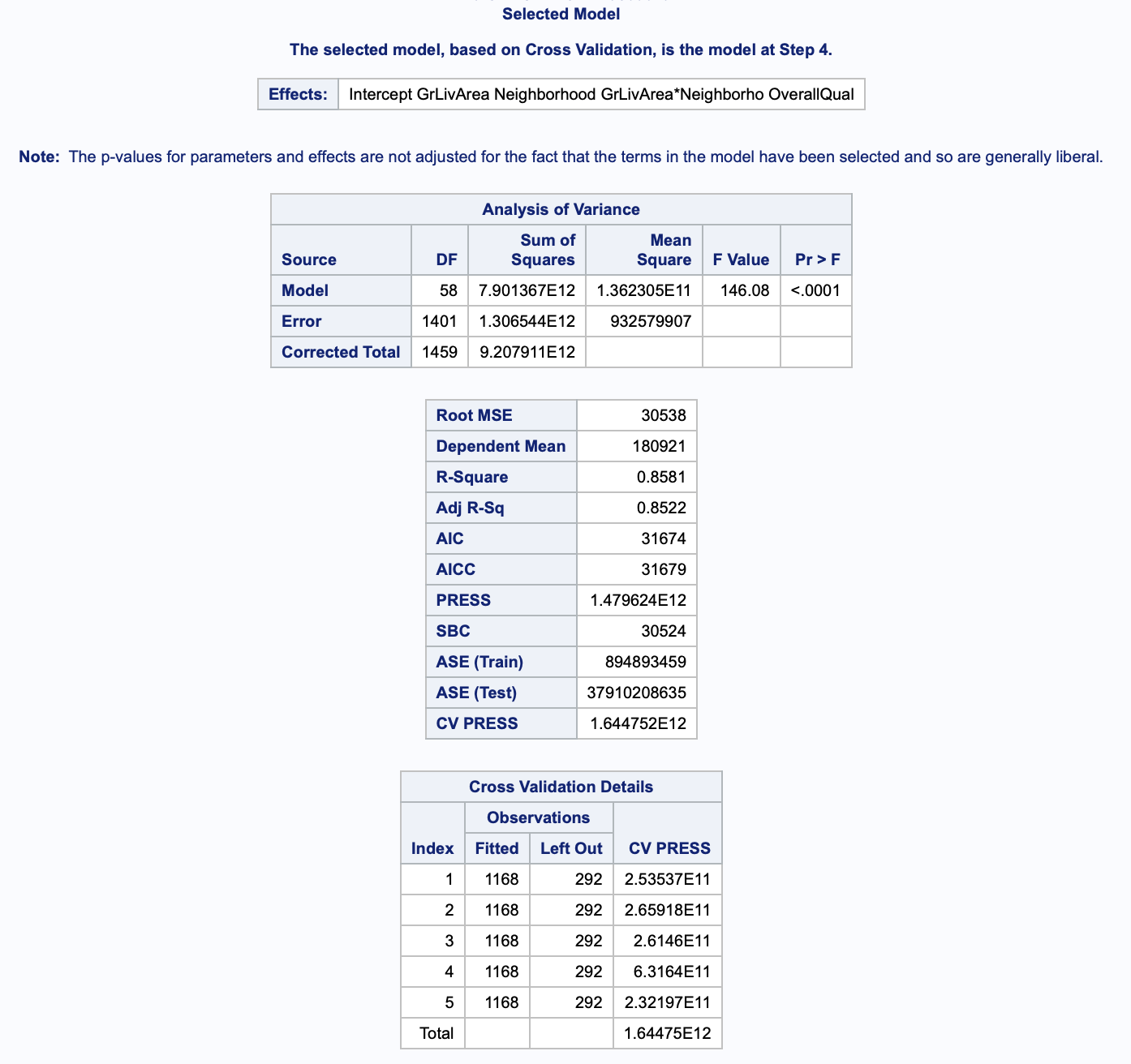
Leveraging the full model to make predictions – we will now attempt to select 4 types of models. All models are linear regression based – but the way in which we will select parameters will differ. In order to simplify things for SAS – we will limit our interaction combinations to a depth of 2 (i.e only interactions of the form A:X or A:B will be considered)

* Forward Selection: this process involves starting with an empty model and sequentially ***adding*** parameters according to the maximum ***positive*** contribution to the selection criteria.
* Backward Selection: this process involves starting with the full model and sequentially ***removing*** parameters according to the maximum ***negative*** contribution to the selection criteria
* Stepwise Selection: A bi-directional approach to variable elimination which allows the gated entry/exit of variables according to include/exclude criteria.
* A stopping condition will be added to reduce over/under fitting of the model.

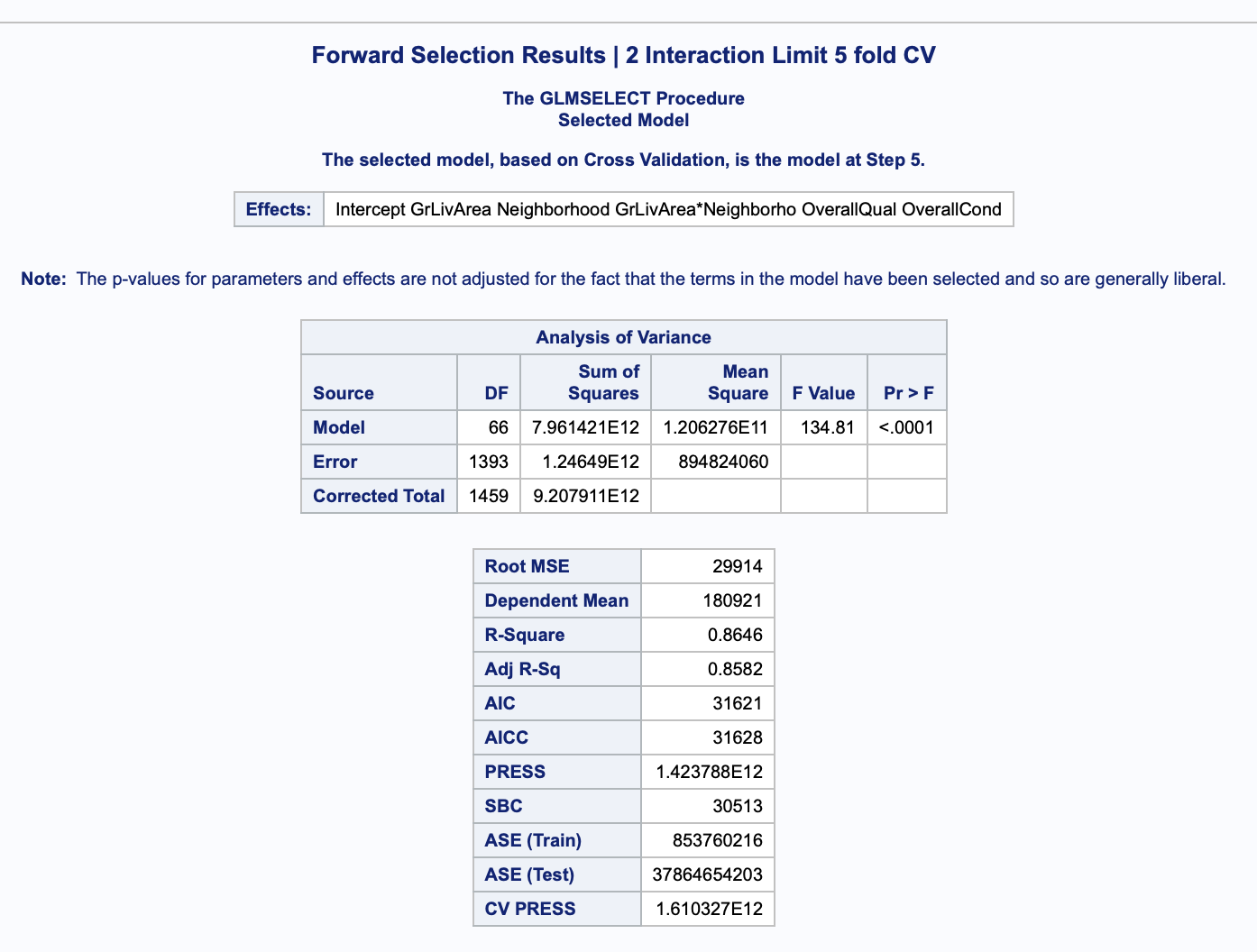
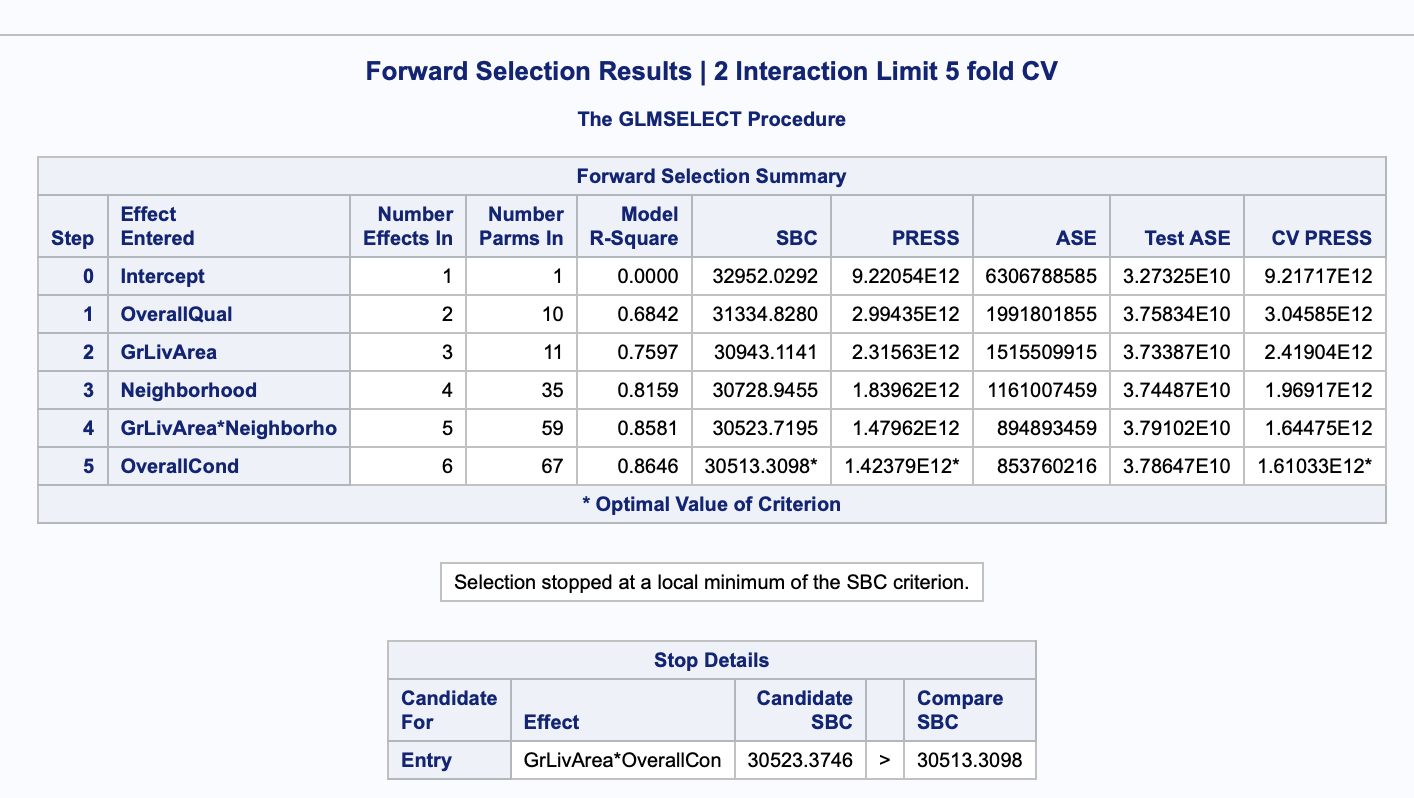
To keep things relatively simple – we will limit our variable search to the following: GrLivArea Neighborhood OverallQuality OverallCondition MSZone – factors that should describe the general characteristics of any house very well – Size (GrLivAre), Location/Social (Neighborhood), Intrinsic Value (Quality/Condition), Type (MSZone)

\*see addendum for detailed analysis of training parameter selection

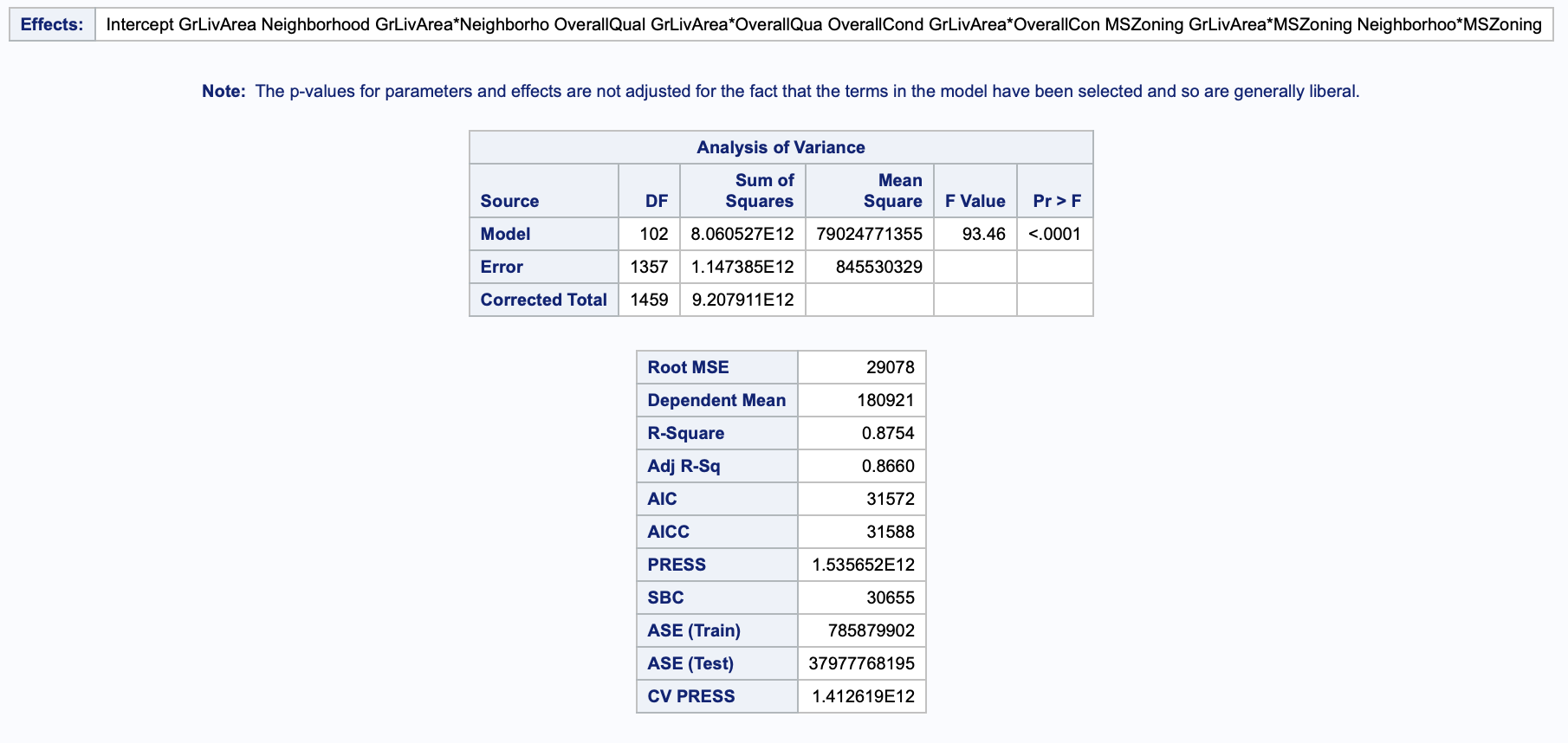
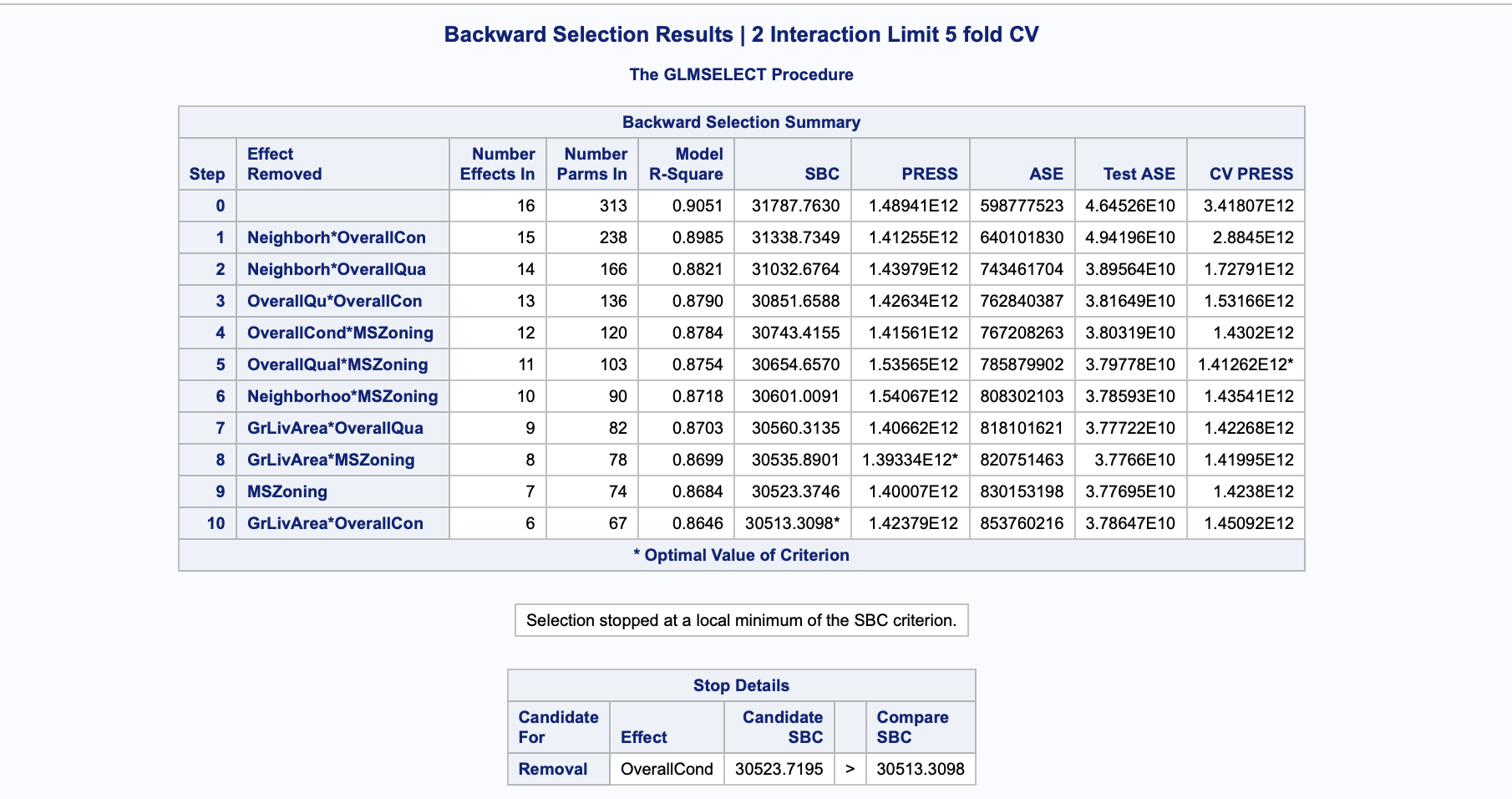
#### Stepwise Results:



#### Forward Selection Results:

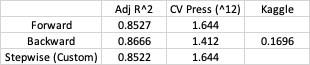
 

#### Backward Selection Results:

#### Discussion:

Ultimately – backward selection was the model selected as our custom model and the one that we submitted to Kaggle for scoring. Backward selection gave the most sophisticated model in terms of interactions but also produced the highest adjusted R squared value. We probably could’ve done better had we used an easier technology to perform creative EDA / model permutation. Results are below:



Explanatory power of each model was relatively similar – though statistical significance of backward model was worst (which is to be expected given the number of interactions considered). Permuting selection criteria for stepwise and adding new features to the selection pool would’ve likely generated a better result\*.

\*HAVING TEAMMATES WOULD’VE ALSO GENERATED A BETTER RESULT ☺